

Programa para análisis de secuencias biológicas

Roberto Reinoso

roberto117343@gmail.com

En este artículo voy a presentar uno de mis programas, que sirve para realizar análisis de secuencias tanto de ADN, ARN como de proteínas.

Este programa esta realizado en Java bajo la licencia de software libre GNU GPL [1]. Disponible el código fuente en la siguiente dirección:

(<https://github.com/roberto117343/biolisis/blob/master/main.java>).

Dicho software realiza el análisis calculando el valor de la longitud de la secuencia introducida y la proporción de nucleotidos o aminoácidos que esta posee.

El funcionamiento del programa es del siguiente modo: una vez seleccionado el tipo de secuencia, para realizar el calculo de la longitud de la misma, este nos devuelve la longitud de la cadena de entrada.

Para calcular las proporciones, el programa, lee el primer carácter de la secuencia introducida. Cuando dicho carácter coincide con uno de los caracteres posibles (estos son las letras correspondientes a los nucleótidos o aminoácidos según el caso) suma 1 a un contador disponible para ese carácter y a continuación pasa al siguiente. Si un carácter no corresponde a ninguno de los incluidos en el programa, este lo cataloga como dato anómalo. Una vez leída toda la secuencia, el programa pasa el valor de cada contador a un porcentaje y lo muestra en pantalla junto con la longitud.

El uso del programa es muy sencillo. Lo primero que debemos hacer es seleccionar el tipo de secuencia que vamos a analizar. Para ello tenemos que introducir 1 para ADN, 2 para ARN y 3 para proteínas.

Una vez hecho esto introducimos la secuencia a analizar. Esta debe ir como una cadena de texto

sin espacios, ni saltos de línea, etc. El programa devuelve primero el tamaño en pares de bases o en aminoácidos (dependiendo de la secuencia introducida) y la proporción de nucleótidos o aminoácidos (según corresponda) en tanto por ciento. Una vez hecho esto el programa nos pregunta si queremos hacer otro análisis.

```
¿Qué tipo de secuencia desea analizar?
```

1. ADN
2. ARN
3. Proteínas

```
1
```

```
Introduce la secuencia a analizar:
```

```
RESULTADOS:
```

```
Tamaño de la secuencia: 35937 pb
```

```
PROPORCIONES
```

```
Proporción de ADENINA: 23.212844700448006 %  
Proporción de CITOSINA: 27.95169324094944 %  
Proporción de GUANINA: 27.2504660934413 %  
Proporción de TIMINA: 21.584995965161255 %  
Proporción de DATOS ANÓMALOS: 0.0 %
```

```
¿Desea analizar otra secuencia? s/n
```

Figura 1. Ejemplo del programa para una secuencia de ADN.

En el ejemplo he usado el genoma completo de un adenovirus [2] y uno de sus productos proteicos [3]. Estas secuencias las podemos encontrar en las bases de datos del NCBI (<https://www.ncbi.nlm.nih.gov/>).

El programa ha sido probado en la consola utilizada por el software Eclipse [4] y en el sistema operativo windows 10. Recomiendo usarlo desde el propio Eclipse, debido a problemas que puede dar el hacerlo fuera de este por introducir una cadena de texto demasiado grande. También mencionar que aunque el programa no este limitado a un tamaño de secuencia determinado, para evitar posibles errores aconsejo usarlo en secuencias inferiores a 50.000 caracteres.

¿Qué tipo de secuencia desea analizar?

1. ADN
2. ARN
3. Proteínas

3

Introduce la secuencia a analizar:

MATPSMMPQWSYMHISGQDASEYLSPLVQFARATETYFSLNNKFRNPTVAPTHDVTTRDQRQLTRFIPVDREDTAYSYKARFTLAVGDNRVLDMASTYFDIRGVLDRGPTFKPYSGTAYNA

RESULTADOS:

Tamaño de la secuencia: 968 aa

PROPORCIONES

A = 7.128099173553719 % R = 4.75206611570248 % N = 7.74793388429752 % D = 6.714876033057851 % C = 0.7231404958677686 %
Q = 4.132231404958677 % E = 5.06198347107438 % G = 6.6115702479338845 % H = 1.5495867768595042 % I = 3.615702479338843 %
L = 7.74793388429752 % K = 4.028925619834711 % M = 2.9958677685950414 % F = 4.648760330578512 % P = 6.095041322314049 %
S = 6.6115702479338845 % T = 6.921487603305785 % W = 1.3429752066115703 % Y = 5.8884297520661155 % V = 5.681818181818182 %

DATOS ANÓMALOS: 0.0 %

¿Desea analizar otra secuencia? s/n

Figura 2. Ejemplo del programa para una secuencia proteica.

A modo de conclusión indicar que el programa está disponible para todo aquel que requiera realizar cálculos en secuencias tanto de nucleótidos como de aminoácidos del tipo antes expuesto. De encontrar algún problema reportadlo y será solucionado lo antes posible.

Referencias

1. GNU General Public License. (2017, 3 de abril). *Wikipedia, La enciclopedia libre*. Fecha de consulta: 03:32, abril 4, 2017 desde https://es.wikipedia.org/w/index.php?title=GNU_General_Public_License&oldid=98087179

2. Human adenovirus 2, complete genome. (s.f.). Desde https://www.ncbi.nlm.nih.gov/nuccore/AC_000007.1

3. hexon [Human adenovirus 2]. (s.f.). Desde <https://www.ncbi.nlm.nih.gov/protein/56160500>

4. Eclipse (software). (2017, 3 de abril). *Wikipedia, La enciclopedia libre*. Fecha de consulta: 03:44, abril 4, 2017 desde [https://es.wikipedia.org/w/index.php?title=Eclipse_\(software\)&oldid=98069545](https://es.wikipedia.org/w/index.php?title=Eclipse_(software)&oldid=98069545)